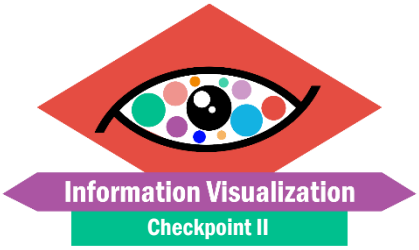

Checkpoint II: Data Cleaning & Processing

Group: G13
Date: 2023/09/25

Initial Dataset

The raw forex price dataset, sourced from Dukascopy, captures price data for seven major currency pairs, encapsulated in OHLC (Open, High, Low, Close) format alongside trading volume. Each entry has a corresponding date and time stamp. The dataset is available in CSV format and consists of the following columns: date, Open, High, Low, Close, and Volume. This dataset offers an in-depth perspective on price fluctuations across different timeframes derived from hourly data. A sample from this dataset has size (305.200, 13) summing all seven datasets and appears as:

	Gmt time	Open	High	Low	Close	Volume	Log_Return	SMA_20	SMA_50	RSI	Upper_Bollinge
0	05.01.2016 23:00:00.000	0.71597	0.71676	0.71590	0.71595	3.114020e+06	-0.000014	0.716901	0.719221	41.257996	0.
1	06.01.2016 00:00:00.000	0.71594	0.71713	0.71584	0.71688	3.493640e+06	0.001298	0.716701	0.718969	49.281314	0.

The economic calendar dataset, procured from FXStreet, chronicles major financial events influencing the forex market. This data contains information such as the event's unique identifier, its start time, name, expected impact, and the likely affected currency or currencies. Each year's data from 2016 to 2023 is saved in distinct CSV files. This dataset has size (58413, 5) and appears as:

	Id	Start	Name	Impact	Currency
0	289f7bf0-e07a-4ae5-be3e-c3099d7d57b0	01/01/2016 00:00:00	New Year's Day	NONE	USD
1	daeff945-2698-4642-82f8-526cd0b207e1	01/01/2016 00:00:00	New Year's Day	NONE	NZD

Selected/Derived Data

The raw forex price dataset underwent a comprehensive analysis, with all columns—date, open, high, low, close, and volume—being retained. Interestingly, despite the vastness of our data, no missing values were found, ensuring that our preliminary analyses would not be skewed. In the context of outliers, while they're often red flags in many datasets, they represent genuine market activities in the forex realm and were, therefore, preserved. To augment our dataset's depth, various derived measures were added: **Log returns -** this represents the logarithmic difference between consecutive close prices, giving a percentage change that accounts for both uptrends and downtrends; **Simple Moving Averages (SMAs)** - the 20-period SMA gives the average of the previous 20 close prices, reflecting short-term trends. Similarly, the 50-period SMA uses the last 50 close prices, offering insights into medium-term movements; **Relative Strength Index (RSI)** - this metric divides the average gain of up periods by the average loss of down periods over a specified interval (typically 14 periods). It then scales this ratio to fit a range of 0-100, helping traders discern potential overbought or oversold market states; **Bollinger Bands** - Originating from the 20-period SMA, the upper band is calculated by adding two standard deviations of price over the period, and the lower band is the 20-period SMA minus two times the standard deviation. These bands adapt to price volatility; **Average True Range (ATR)** - this gauges volatility by determining the greater of the following: the current high minus the current low, the absolute value of the current high minus the previous close, or the absolute value of the current low minus the previous close. A 14-period average of these values gives the ATR. Rows with "NaN" values, primarily stemming from the initial periods of these derived measures, were subsequently removed to maintain data consistency.

Shifting to the economic calendar dataset, which sheds light on significant financial events, it was concatenated from annual segments spanning 2016 to 2023. Every original column was retained, and through merging the annual fragments, redundancy was curbed by discarding duplicate 'Id' entries. Impressively, this dataset too bore no missing values. Nonetheless, to ensure data integrity, rows with "NaN" values, mainly from preliminary periods, were excluded.

Data Abstraction

Both the forex price dataset from Dukascopy and the economic calendar data from FXStreet are tabular datasets. They are presented in structured tables, with each row representing an item (a data entry) and each column representing an attribute (a specific type of information).

The dataset comprises attributes essential for forex analysis. 'Open' represents the starting price of a time period, while 'Close' indicates its end. 'High' and 'Low' record the maximum and minimum prices, respectively, during that period. 'Volume' measures total trading activity. In the economic calendar data, 'Id' uniquely labels each news event. 'Name' offers a brief description, 'Impact' gauges the potential market influence, and 'Currency' pinpoints which currencies might be affected. These elements, combined, give a holistic overview of price actions and market-influencing news.

- Open, High, Low, Close, SMAs, RSI, ATR = (Ratio, Sequential, Non-Hierarchical)
- Volume = (Ratio, Divergent, Non-Hierarchical)
- Log Returns, Bollinger Bands = (Continuous, Sequential, Non-Hierarchical)
- Date, Start = (Continuous, Cyclical, Non-Hierarchical)
- Impact = (Ordinal, Sequential, Non-Hierarchical)
- Id, Name, Currency = (Categorical, Sequential, Non-Hierarchical)

Data Processing

In the data processing phase, we utilized Jupyter Notebooks. Within this platform, we managed all preparatory tasks, from addressing potential missing values and adding derived measures to removing "NaN" values. The interactive nature of Jupyter Notebooks facilitated real-time visualization and immediate feedback, ensuring our datasets were optimally prepared for analysis.

Mapping (Data sample/Questions)

1. Interactive Exploration of Price Movements: The OHLC prices and date attributes facilitate the creation of a candlestick chart, allowing users to dissect currency-specific trends across time frames and identify classic price patterns. **2.** Price Movement Correlation with News: By merging the date attribute from both the forex price and economic calendar datasets, we can gauge the impact of key news events on price variations, offering insights into the market's reactions and broader shifts. **3.** News Volume in each Country: Utilizing the 'Country' and 'Impact' attributes from our news dataset, we can craft a choropleth map. This visual will display both the volume and significance of news coverage across countries. **4.** Currency Pairs Correlation Matrix: Through the OHLC prices of various currency pairs, we can generate a correlation matrix. Presented as a heatmap, users can discern interrelations between different currency pairs. **5.** Influence of Technical Indicators on Price Behaviour: Incorporating derived measures like the Simple Moving Averages, RSI, and Bollinger Bands with price data lets users juxtapose technical indicators with price movements, pinpointing potential trading signals or anomalies.